DOCUMENT RESUME

ED 322 220 TM 015 505

AUTHOR	Gugel, John F.
TITLE	A Normalized Direct Approach for Estimating the Parameters of the Normal Ogive Three-Parameter Model for Ability Tests.
INSTITUTION	National Center for Education Statistics (ED), Washington, DC.
PUB DATE	Jul 90
NOTE	41p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Ability Identification; Aptitude Tests; Chi Square; Comparative Analysis; *Computer Assisted Testing; *Estimation (Mathematics); Goodness of Fit; *Mathematical Models; *Multiple Choice Tests; Regression (Statistics)
IDENTIFIERS	*Ability Estimates; LOGIST Estimation Procedures; Normalization Regression Estimation; *Normal Ogive Models; Three Parameter Model

ABSTRACT

A new method for estimating the parameters of the normal ogive three-parameter model for multiple-choice test items--the normalized direct (NDIR) procedure--is examined. The procedure is compared to a more commonly used estimation procedure, Lord's LOGIST, using computer simulations. The NDIR procedure uses the normalized (mid-percentile) "z" score as the ability score as opposed to the conventional raw score (linear "z" score), maximum likelihood, or Bayesian modal ability score. Thus, it is not necessary to use an iterative procedure for estimating the person parameter (ability); corrections for scale errors can be made before the item function fitting is complete. The item function fitting is accomplished using a minimum chi square procedure. Input to the chi square procedure includes biserial correlations corrected for guessing and attenuation. The attenuation correction uses a reliability index based on the item set information index. Normalized "z" scores corrected for attenuation using the KR20 index are also inputted to the chi square procedure. The normalized direct procedure was more accurate and required considerably less computer processing time than did LOGIST. One figure and six data tables are included. (Author/TJH)

A Normalized Direct Approach for Estimating The Parameters of the Normal Ogive Three-Parameter Model For Ability Tests

John F. Gugel, Ph.D.
National Center for Education Statistics
U.S. Department of Education

June 1990

July 1990

This document reflects the views of the author and not necessarily those of the National Center for Education Statistics.

The author would like to express his graditude to a number of people who made helpful comments along the way. These include Dr. Lois Northrup, Dr. Mary Anne Nestor, and Mr. John D.Kraft of the Office of Personnel Management; Dr. Charles E. Johnson of the University of Maryland; and Raul Garza of the National Center of Education Statistics.

The author may be contacted at the following address and number:

John F. Gugel
7227 Tyler Ave.
Falls Church, Va. 22042
Tel: 703 573 4352

ABSTRACT

A new method for estimating the parameters of the normal ogive three-parameter model for multiple choice test items, the normalized direct (NDIR) procedure, is examined and compared to a more commonly-used estimation procedure, Lord's LOGIST, using computer simulations.

The NDIR procedure uses the normalized (mid percentile) z score as the ability score as opposed to the conventional raw score (linear z score), maximum likelihood, or Bayesian modal ability score. Thus it is not necessary to use an iterative procedure for estimating the person parameter (ability); corrections for scale errors can be made before the item function fitting is completed.

The item function fitting is accomplished using a minimum chi square procedure. Input to the chi square procedure includes biserial correlations corrected for guessing and attenuation. The attenuation correction uses a reliability index based on the item set information index. Normalized z scores corrected for attenuation using the KR20 index are also inputted to the chi square procedure.

The normalized direct procedure was more accurate than LOGIST and required considerably less computer cpu time.

The normalized direct (NDIR) procedure is a new method for estimating the parameters of the normal ogive three-parameter model for multiple choice test items. Unlike Lord's LOGIST program which uses iterative procedures to compute a person's ability parameter from item score pattern likelihoods based on rough initial parameter estimate values, the NDIR procedure uses normalized z scores (referred to as z' in this paper) as person parameter estimates.

In LOGIST, the initial item parameters are usually based upon linear z transformation of raw scores. In the NDIR procedure, the normalization is based upon the initial raw score distribution. It computes the person parameters directly from the raw score frequencies using the midpercentile z of the raw scores (z') instead of the raw scores themselves. Thus scaling errors due to skewness may be greatly reduced, i.e., it converts distributions to approximately normal form "... thus precluding the errors in correlation matrices which arise from disparity of shape" (Carroll, 1961, p. 359).

As will be demonstrated, the true scale versus estimated scale errors do not change sign as readily with the midpercentile z scores when guessing increases as with linear z scores. Thus corrections for attenuation (error in the scale) can be made

before the item function fitting is completed. That is, the estimated ability scale can be brought to more nearly approximate the true ability scale. It is not necessary to estimate the guessing parameter from the number of item alternatives, and there is virtually no limit on the number of examinees that may be processed at one time.

The Model

The three-parameter model is a mathematical function that yields the probability value of examinees who would obtain an item score of 1 (get the item right) at a given ability score scale point. The ability (θ) scale usually ranges from minus infinity to plus infinity, but it is rare for ability distributions to fall outside the range -3.0 to + 3.0. The function has an "a" parameter for item discrimination, a "b" parameter for the item difficulty (scale point of peak item functioning), and a "c" parameter for item guessing (the probability that those who do not know the item will respond correctly by guessing).

Currently there are two forms of the three-parameter function: (1) the normal ogive and (2) the logistic model (often called the Birnbaum three-parameter model). The normal ogive version uses the unit standard normal curve. The probability of

examinees getting an item score of 1 (correct answer) at an ability point, θ, before guessing, is the area, p(θ), under the normal curve from the base scale point to the upper end. The plot of the function (normal ogive or logistic) is called an item characteristic curve (ICC). An ICC is a curve plotted on the points representing the probabilities of examines getting the item right at various θ level points. The actual normal distribution formula is

$$p(\theta) = 1/(2\pi)^{1/2} \int_{-a(\theta-b)}^{\infty} e^{-t^2/2}\, dt$$

and the logistic function approximates the normal ogive function using the formula

$$p(\theta) = 1/\{1 + e^{-1.7a(\theta - b)}\}$$

According to Lord and Novick (1968, p. 399) the logistic function differs from the normal by less than .01 for all θ. Many psychometricians prefer the logistic form because, mathematically, it is simpler.

Illustration of the Three-Parameter Normal Ogive Model

The three-parameter normal ogive model is depicted in

Figure 1. The assumption is made that the hypothetical

[Insert Figure 1 about here]

distributions underlying the item scores of 0 or 1 (Distribution 2 on the vertical axis of Figure 1) and that of the ability trait (test trait denoted by Distribution 1 on the horizontal axis of Figure 1) are normal. The test is assumed to be measuring one trait (unifactor) and the response to one item not dependent on any other item (locally independent). These two axes with the distribution density axis perpendicular to them represent a bivariate surface, a three dimensional figure which has been collapsed on the ordinate axis and then on the abscissa. The regression line has as its slope $\rho_{I\theta}$, the correlation (biserial corrected for guessing and attenuation) between the θ trait and the item. The unconditional probability of a correct answer to an item (average difficulty or p-value) is represented as the shaded area under the θ_I unit standard normal distribution (Distribution 2) of Figure 1. Figure 1 might represent an item for which the unconditional probability of a correct response is .85.

P Value for an Item at a θ Point

The fixed point on the θ_I scale representing the shaded area point of cut in Distribution 2 is defined as γ. A line extended across the bivariate surface at that fixed point and parallel to

the θ-axis will intersect all possible θ slices (conditional θ_I distributions) of the bivariate surface. The probability, $p(\theta)$, of passing the item at any value θ is represented by the portion of the θ_I distribution at that point γ to the upper end (see distributions 3 and 4 in Figure 1). Distributions 3 and 4 represent slices of the bivariate surface at values of θ (conditional θ distributions). Distribution 3 represents the slice of the bivariate at $\theta = b$ and Distribution 4 represents a slice of the bivariate surface at any given point of θ.

The b Parameter

The difficulty parameter b for a specified item is defined as that value of θ where $p(\theta) = .50$. The b value is, therefore, on the same scale as the ability scores. It is also the θ value corresponding to the item score mean (P value). In addition to being the slice of the bivariate surface at $\theta = b$, it is the θ coordinate for the point of intersection of the line $y = \gamma$ and the regression line, or

$$b = \gamma/\rho_{I\theta} \tag{1}$$

The a Parameter

The point where the line $y = \gamma$ intersects Distribution 4 of

Figure 1 would have the value $\rho_{I\theta}(\theta - b)$ on the θ_I scale. To convert this into a value on the Distribution 4 scale, it is necessary to divide it by the standard deviation of Distribution 4. Its standard deviation is

$$(1 - \rho_{I\theta}^2)^{1/2}$$

When γ is taken as a point on the baseline of Distribution 4, the following is true:

$$\gamma = (b - \theta)\, \rho_{I\theta}/(1 - \rho_{I\theta}^2)^{1/2}$$

As noted in Lord and Novick (1968, p. 378):

$$a = \rho_{I\theta}/(1.0 - \rho_{I\theta}^2)^{1/2} \qquad (2)$$

Equation 2 represents the item-ability scale correlation coefficient divided by the regression line standard deviation. The γ point then becomes $-a(\theta - b)$ and $p(\theta)$ is the area from that value to the upper end (shaded in Distribution 4).

A unit standard normal table can be consulted to obtain the area. The programs developed for this study use the normal ogive

form of the model. The IBM (1968) subroutines are easy to use and are accurate to more than three decimal places, so it is unnecessary to use the logistic form to approximate the normal ogive model.

The c Parameter

When guessing is involved, the proportion of examinees who score '1' on the item exceeds the expected P-value (P) in the total group (and $p(\theta)$ at a given θ point). The probability that examinees who do not know an item get the item correct through successful guessing is c. That is, when there is guessing, $c(1 - P)$ more examinees would get the item right and the total would be $P + c(1 - P)$ or, conditionally, $p(\theta) + c(1 - p(\theta))$ at a given point. Taking P' to be the guessing-present total, $q(\theta) = 1 - p(\theta)$, and $p'(\theta)$ the θ point probabilities of getting the item right, the following formulae can then be written:

$$P = P' - c(1 - P) \qquad (3)$$

or

$$p'(\theta) = p(\theta) + cq(\theta) \qquad (4)$$

and

$$p(\theta) = p'(\theta) - cq(\theta) \tag{5}$$

Thus $p(\theta)$ in Equation 5 can be obtained by removing from $p'(\theta)$ the proportion due to guessing. The c parameter is also the lower asymptote of the ICC.

The c parameter is not an explanation of how examinees guess. It simply acknowledges that they guess. As the ability level of examinees increases, the number of those that guess declines since it can be assumed that people who guess come from among those who do not know the answer. The parameter c thus may be described as a noise parameter added to the two-parameter model.

Consequences of Guessing

As will be shown, guessing tends to make normal distributions of raw scores appear non-normal. When guessing is ignored severe inaccuracies can result in computations using the item score or the total raw score. As noted, $\rho_{I\theta}$ is needed in Equations 1 and 2 to compute both the a and b parameters. γ is computed from the proportion passing the item with guessing absent and is used in Equation 1 to compute the b parameter. Correcting for guessing is important (see Guilford, 1954, p. 421). That is, when .10 of a population know an item, and of the remaining .90 only .20 get it correct through successful guessing, the observed P-value (P') is

.28 (2.8 times as high as it should be).

Easy items, however, are not affected as severely by guessing as are the hard ones. For example, if those who know the item constitute a proportion of .90 and, .20 of the remaining .10 is added to that proportion, the observed P' would be .92 or only 1.02 times as high as it should be (instead of 2.8 as observed above on the other end of the scale).

The population value $\rho_{I\theta}$ is also severely attenuated by guessing (see e.g., Urry, 1977; Ashler, 1979). The value 1/k, where k is the number of item alternatives, traditionally has been used as if it were an efficient estimator of the c parameter. However, evidence cited in Rowley and Traub (1977) indicates that equations using 1/k are generally not appropriate for unspeeded tests.

The Ability Score Problem

Procedures like LOGIST that estimate the item parameters require estimates of the person parameter θ. Only two ways of estimating θ have been proposed that do not require fallible starting estimates of the item parameters. These are the z and z' methods. All other techniques use probilities of pattern scores. How close to the true θ values do z and z' come? Can appropriate adjustments be done to make meaningful improvements? Table 1

was computed to investigate these questions.

Table 1 was created from a larger matrix that has chosen θ values from θ = -4.5 to θ = 4.5 with increments of .1. At each θ step the likelihood of each score pattern of 0's and 1's was computed using the normal ogive three-parameter model and the item parameters given on page 22. The likelihood value (combined product of the p'(θ) and q'(θ) values) of the vector was multiplied by the normal distribution frequency at the θ point. Thus a θ by raw score matrix was created. The mean θ value was computed at each raw score, as indicated in Table 1.

[Insert Table 1 about here]

Table 1 is based on a 40-item test with a known distribution of item parameters and a known unit standard normal θ distribution. The a and b parameters are those used for later simulations and, as mentioned, are on page 22. Three levels of assumed guessing are compared. For each possible raw score (X) the expected frequency was computed, the θ mean of all 0,1 response patterns giving a particular expected X, z score from the expected X, and the z' from the expected cumulative frequency at the middle of the interval.

In Table 1 the first distribution is reported for the set of 40 items with no guessing; the second is for the same item set with

all guessing parameters set to .10; and the third is the same set with guessing set to .20.

Raw scores (X) are given in column 1, Distribution 1 (no guessing) in columns 2-5, Distribution 2 (guessing = .10) in columns 6-9, and Distribution 3 (guessing = .20) in columns 10-13. The person parameter is normally distributed. The likelihood probability response vectors associated with each raw score are generated using the normal ability values and the established item a, b, and c parameters. Thus it is possible to construct an expected frequency distribution for any sample and also to observe the internal distribution of θ's for any raw score level. For example, for Distribution 1 (no guessing) a raw score of 25 has a proportion point .039 (column 2), which means that in a sample of 1,000 subjects 39 people are expected to have a score of 25. The average θ of these 39 persons is expected to be .55 (column 3). the raw score of 25 corresponds to a linear z of .61 (column 4) and a z' of .57 (column 5). The mean raw score of the 40 items is expected to be 20 and the standard deviation 8.24. The table thus provides a theoretical picture of the relationship of the raw score distribution and its transformations to the underlying θ distribution.

Table 2 presents the discrepancies between the z and z'

values and the θ mean for each score level.

[Insert Table 2 about here]

Note that the $z' - \theta$ mean discrepancies in each set are less than the corresponding $z - \theta$ mean discrepancies at the high frequency points of the scale, and the discrepancies for z' (attenuation) tend to be all in the same direction (away from the middle of the distribution). The direction of the linear z errors, on the other hand, tends to vary, especially as guessing increases.

Note in Table 2 that when $c = 0$ the discrepancies are positive over the range 0 to 4 then negative in the range 5 to 19 and negative again at 37 and above. When $c = .10$ the discrepancies are negative for the ranges 0 to 20 and 37 to 40, etc. It is not always possible to tell where the discrepancies will be negative or positive. Since the extent of guessing varies from item to item and is not known, it is extremely difficult to make adequate corrections for the error caused by the linear z. This may introduce fluctuating bias (as well as some possible random error) into initial pattern probabilities and ability score estimates that may never be overcome in later parameterization stages or cycles. That is, as noted in Baker (1987), "there is a common thread that runs through all the available estimation methods. As a result they share the

limitations, such as the need for good initial estimators and well-conditioned matrices, of the Newton-Raphson technique" (p. 138).

The Normalized Direct Procedure

Note in Table 1 that the shape of the z' distribution is always normal in form. The shape of the initial ability distribution does not shift considerably when different amounts of examinee guessing occurs as does the z distribution. Yet little research has been done on the applicability of the z' scale to item parameterization, in spite of the fact that Johnson's (1951) U-L (Upper-Lower 27%) index (influenced by Kelly's (1939) frequency work) has been in use for many years.

If the item-score biserial is computed using z' scores instead of raw scores, both the item and raw score distributions are normal in form. This happens because the biserial already requires the item score mean be converted to unit standard normal values. Hence some serious problems with estimating the biserial are avoided; as noted in Carroll (1961, p. 359), the transformation of distributions to an approximately normal form is a common way to improve estimates of correlations. Guilford and Fruchter (1978, pp. 307, 478-484), note that grave errors such as biserials in excess of 1.0 can result when the raw score

distribution is left in a very skewed form.

As noted in Schmidt (1977), an attenuation correction can be applied to $\rho_{I\theta}$ before the a and b parameters are computed. Evidence presented in Tables 1 and 2 suggests that z' errors can be appreciably reduced using a reliability coefficient correction similar to the one suggested by Schmidt since the θ - z' errors tend to be of the same sign whether or not there is guessing. When a successful correction is made before the function fitting process is completed, less error would be expected in the item parameter estimates. This would be expected to show up when computer simulations are done with estimated values compared against known values and the LOGIST procedure.

Steps of the Procedure

The procedure entails: (a) Scoring of the test and storage of item-test data, (b) Choosing and using provisional (trial) guessing values for steps to follow and using guessing corrected item-raw score biserials to estimate a and b parameters, (c) Fitting trial item characteristic curves to the data item characteristic curves using a chi square procedure to determine the best c parameter estimate and the corresponding a and b parameter estimates, (d) Using the a, b and c estimates to compute a correction for attenuation to $\rho_{I\theta}$ which is used with a

correction to z' based on KR20 as steps 2 and 3 are repeated to yield final estimates.

Scoring and Storage

Examinees' 0,1 item scores are summed to obtain raw scores. An item by score matrix of frequencies is created and stored for the two fitting cycles. Then, from the total group frequency distribution of all examinees, z' scores are computed from each point on the test score distribution. All item statistics are computed with all the other item scores included in the total score but not the score of the item itself. The z' scores (θ estimates) are also stored for first and second cycle use. Two IBM (1968) subroutines (NDTRI and INGRAT) are used to convert areas under the normal curve to baseline values and vice versa.

Using Trial Guessing Values and Computing First Parameters

The stored z' values and frequencies are used to compute the item-raw score biserial correlation coefficients. For each item, provisional (trial) c values between .000 and .320 are used. This is done in the following manner. First, computations (which will be explained) are made using c = 0, then .04. .08, .12, .16, .20, .24, .28, and .32. Then, as determined by the computations, smaller steps are used between two of these values until the computational criteria are satisfied. The a parameter comes from

an item-raw score biserial corrected for guessing using the respective trial c value. Equation 3 is used to remove the effect of guessing in the P-value which is then converted to γ and ϕ values for the biserial and also the b parameter.

When the biserial with guessing present is computed, it is assumed that examinees who guess (as mentioned earlier) come from among those who do not know the item. Their mean true ability is the same as that of those who do not know the answer. It is assumed that for each examinee, success in guessing an item is independent of success on any other item. Then the guessing-corrected biserial is

$$r_{bc} = p'(\bar{X}' - \bar{X})/[(1 - c)(\sigma_X \phi)] \qquad (6)$$

where $\bar{X}$ is the mean z' score of all who took the test, $\bar{X}'$ is the mean z' score of the observed number of examinees obtaining the correct answer, σ_X is the total group standard deviation, and ϕ is the height of the ordinate (corresponding to the area under the normal curve for the corrected proportion obtaining the correct answer). In Figure 1, Distribution 2, ϕ is the height of the curve above γ on the base. It is obtained using Equation 5 and the unit standard normal tables (See also Ashler, 1979). The

score means in Equation 6 are assumed to be continuous. The z' score conversion converts the score distribution to a normal form. Thus both the item and test score distribution behave as if they are continuous. The biserial should have a continuum on both the item score distribution and the test score distribution. (see Lord, 1980, p. 33).

Fitting the ICC to the Data

A minumum chi square estimation routine is now employed that uses the item-by-z' score matrix of frequencies, each trial c value and the corresponding a and b values. This chi square routine has been used by a number of psychometricians, especially Urry (1976, p. 17). The chi square (sum of squares fit or SSFit) formula may be given in the form:

$$\text{SSFit} = \sum_{j=0}^{m-1}[n'_j - t_j p'_j(z')]^2/[t_j p'(z')q'(z')] \qquad (7)$$

where j represents the test score, m the total number of items, n_j' the number of cases obtaining the right answer on the item at the score, j, t, the number of people at the score j, and p'(z') is the expected (computed) proportion passing the item at the ability level.

A fitting sequence goes as follows. A trial value of the c

parameter is chosen. Equation 5 is used to compute P, which in turn is used to get γ and ϕ. Then Equation 6 is used to compute r_{bc}, which in turn is used in Equations 1 and 2 to get the b and a parameter estimates, respectively. Next, with a and b values corresponding to the trial c value, Equation 7 is used. However, first a p'(z') must be calculated for each z' using -a(z' - b), its corresponding probability from the normal table, and Equation 4. Then the SSFit value is compute using Equation 7. This process is repeated for <u>each</u> trial c value. The a, b, and c values yielding the smallest SSFit are chosen as the best estimates.

Repeat of Steps 2 and 3

Since $\rho_{I\theta}$ requires true parameter values, it is necessary to first get estimates of the three parameters in the first ICC fitting cycle prior to any correcting. Then in the second ICC fitting, the z' values entered into the SSFit subroutine are multiplied by the square root of KR20 to reduce the discrepancies shown in Table 2; r_{bc} is corrected for attenuation to produce an estimate of $\rho_{I\theta}$ using the equation:

$$\rho_{I\theta} = r_{bc}/(r'_{\theta z})^{1/2} \qquad (8)$$

The attenuation correction, $r'_{\theta z}$, is based on the information function at the item b value, and will be explained in the next section.

The Scale Variance

A variance for the estimates of ability for (0,1) scoring at an ability point θ is given in Lord and Novick (1968, p. 453-457) as the inverse of the information function, $I(\theta)$. Lord and Novick's upper bound $I(\theta)$ may be written as

$$I(\theta) = \sum_{i=1}^{m} \lim_{\Delta\theta \to 0} [(\Delta p'(\theta)/\Delta\theta)^2 / (p'(\theta)q'(\theta)) \qquad (9)$$

That is, information at an ability point is equal to the sum, across all m test items, of the squared changes in the proportion passing the item at the ability point (squared derivative of $p'(\theta)$ divided by the respective item variance at the ability point). (The $\Delta\theta$ slices used are $(\theta - .005)$ and $(\theta + .005)$.)

Point Reliability Formula

When the reciprocal of Equation 9 is entered into the equation for the usual reliability coefficient, the θ-point reliability follows. That is,

$$r'_{\theta z} = 1/\{1+1/(\sum_{i=1}^{m} \lim_{\Delta\theta \to 0} [(\Delta p'(\theta)/\Delta\theta)^2/(p'(\theta)q'(\theta))]\} \qquad (10)$$

Since the b parameter is also the regression point corresponding to the item P-value (item score mean), the point reliability provides a reliability index at the θ item difficulty point. Evidence suggests that it is a better index for correcting the item biserial for attenuation than the KR20 suggested by Schmidt (1977). Therefore, NDIR corrects the z' values through multiplication by the square root of the KR20 reliability index before entry into the SSFit equation.

Computer Evaluation of the Procedure

Computer simulations were used to evaluate the item parameter estimate accuracy, since such controlled conditions allow estimation results to be compared to starting (known) conditions. Research-determined item parameters were fed into an item vector response generator. For the item response generator, a random value of θ was obtained for each simulated examinee by summing twelve random numbers in the range .00 to 1.00 from a uniform (rectangular) distribution and subtracting 6 from the sum. According to Morrison (1967, pp. 4-9), the resulting number is normally distributed with a mean of zero and a standard deviation of one. A vector of item responses was then created using this θ, the chosen (known) item parameters, and $p'(\theta)$. This was done by computing the $p'(\theta)$ value for each item. For each item, $-a(\theta-b)$

(Distribution 4 baseline value in Figure 1) was computed and converted to $p(\theta)$ (shaded area of Distribution 4 in Figure 1) using PVAL (adjusted form of the IBM (1968) INGRAT routine); $p(\theta)$ in turn was converted to $p'(\theta)$ using Equation 4 (p. 13).
An equivalent method would be to obtain, from the unit standard normal table, the area under the curve from z (when $z = -a(\theta-b)$) to the upper end. Then for each item a different number between 0.0 and 1.0 can be drawn from the uniform distribution. If the number drawn for the item is larger then $1.0 - p'(\theta)$, the item is scored "1"; otherwise it is scored "0".

Item Parameters Selected for Simulation

Psychometricians generally prefer item difficulties from -2.0 to 2.0 because this difficulty range covers most ranges involved in testing. For this study the b-values were spread in equal intervals (steps) across that range.

The a, b, and c parameters of items to be simulated in the study were:

(1) .60 -2.00 .12; (2) .60 -1.56 .05; (3) .60 -1.11 .07;
(4) .60 -.67 .11; (5) .60 -.22 .00; (6) .60 .22 .19;
(7) .60 .67 .01; (8) .60 1.11 .17; (9) .60 1.56 .24;
(10) .60 2.00 .23; (11) 1.07 -2.00 .01; (12) 1.07 -1.56 .09;
(13) 1.07 -1.11 .02; (14) 1.07 -.67 .10; (15) 1.07 -.22 .05;

(16) 1.07 .22 .12; (17) 1.07 .67 .14; (18) 1.07 1.11 .08;

(19) 1.07 1.56 .08; (20) 1.07 2.00 .13; (21) 1.53 -2.00 .15;

(22) 1.53 -1.56 .11; (23) 1.53 -1.11 .01; (24) 1.53 -.67 .10;

(25) 1.53 -.22 .13; (26) 1.53 .22 .14; (27) 1.53 .67 .14;

(28) 1.53 1.11 .15; (29) 1.53 1.56 .23; (30) 1.53 2.00 .11;

(31) 2.00 -2.00 .14; (32) 2.00 -1.56 .13; (33) 2.00 -1.11 .07;

(34) 2.00 -.67 .07; (35) 2.00 -.22 .11 (36) 2.00 .22 .02;

(37) 2.00 .67 .11; (38) 2.00 1.11 .07; (39) 2.00 1.56 .16;

(40) 2.00 2.00 .03

This 40-item set is repeated to obtain the 80-item set.

Item and Subject Sample Size

A normal ogive generator (see Gugel, 1988, Appendix 2) was used to generate three samples of 40- and 80-item tests each in sets of 1,000, 2,000, and 3,000 cases for a total of 18 sets. The two programs were run on each set -- NDIR and LOGIST. LOGIST was allowed 40 stages for convergence with number of item alternatives set to 4.

The Hulin ICC Recovery Measure

After the item response vectors are entered through the item parameter estimation programs, accuracy of results can be ascertained through comparison with the starting parameter data. Hulin et al. (1982) have provided a very good way of accomplishing

this. Since most measurement applications involve $p'(\theta)$ in the θ range -3.0 to +3.0, they provide the following item root mean square error index (RMSE) for each item:

$$\text{RMSE} = \{\sum_{j=1}^{31} [p'(\theta_j) - p''(\theta_j)]^2/31\}^{1/2} \qquad (11)$$

where j is the number of increments, $p'(\theta_j)$ is computed from the item parameters going into the generator, and $p''(\theta_j)$ from the parameter estimates coming out of an estimation program (i.e. NDIR or LOGIST). The 31 θ_j increments are each .2 (i.e., the θ values are -3.0, -2.8, ... +2.8, +3.0).

Hulin et al. (1982) compute a RMSE for each item and then report a RMSE mean and standard deviation for the item sets. Since the most important θ ranges are used, this index is a more meaningful item parameter accuracy measure than the earlier method which will be discussed later.

The estimation programs were compared using the Hulin index (1982) and independent parameter recoveries.

The Generator Output

The normal ogive vector generator (Program NOVGEN, Gugel, 1987, Appendix 2) was run with the item a, b, and c values given. Starting seven digit generator seed numbers were chosen randomly.

ICC Recovery - Hulin Index

As noted, the Hulin index is an ICC recovery measure. A total of 31 θ points in steps of .2 between -3 and +3 were used. At each point conditional probabilities were computed using the item parameters entered into the item response generator at the start and the corresponding estimated values from NDIR and LOGIST. Each item RMSE was based on the discrepancies between the 31 ICC points. Then a mean and standard deviation based on the item RMSE values in the test was computed. Table 3 contains the mean RMSE values and their standard deviations for each sample and each parameterization method in this study.

[Insert Table 3 about here]

LOGIST had a lower Hulin index in Table 3 in only one of the 18 samples - the first (40 items and 1,000 cases). At 40 items an increase in number of cases seemed to reduce the mean error faster in NDIR than LOGIST (.030 to .023 as opposed to .032 to .031). At 80 items LOGIST dropped from .033 to .023 while NDIR dropped from .027 to .019.

Hulin et al. (1982) reported mean item root mean square errors for LOGIST of .048 at 30 items and 1,000 cases; .043 at 30 items and 2,000 cases; .038 at 60 items and 1,000 cases; and .032 at 60 items and 2,000 cases. The 1,000 and 2,000 case 40-item

means of .032 and .031 for LOGIST in Table 3 appear to be slightly smaller than what they found. They also found magnitude of errors less as number of items increased.

Individual Parameter Recoveries

Item set root mean square errors ($RMSE_p$'s) were computed for the individual item paramerters using the equation:

$$RMSE_p = [\sum_{i=1}^{m}(k - e)^2/m]^{1/2} \tag{12}$$

These are given in Table 4. At 40 items all the cell error means

[Insert Table 4 about here]

are lower for NDIR on all three parameters. At 80 items they are about the same on the a parameter but lower on both the b and c parameters.

Correlations of Recovered Item Paramaeters

Table 5 provides the correlations between the starting parameter values and the corresponding estimated parameters from each technique. The mean correlation for the known versus estimated a parameter was .948 for NDIR and .950 for LOGIST; .994

[Insert Table 5 about here]

versus .992 for the b parameter; and .608 versus .514 for the c

parameter. On the c parameter, unlike the other two, NDIR was best on all nine cell means.

Run Time

Computer cpu times are given in Table 6. The differences in cpu time between the two programs were considerable. NDIR required about 8% of the cpu time of LOGIST.

[insert Table 6 about here]

Conclusions

The normalized direct procedure fares quite well when compared with LOGIST. It differs from LOGIST in that it not only estimates examinee ability scores directly from total raw score frequency distributions using midpercentile z' scores but also applies a correction for measurement scale attenuation to $\rho_{I\theta}$ before the final a, b, c parameters are computed. The correction is based on a reliability coefficient computed from the test information function using first cycle parameters that are corrected only for guessing. This correction process helps correct for the fact that z' is not a perfect replacement for θ.

The findings of this study suggest that LOGIST with more complicated and more costly θ routines may not be doing as good a job at estimating parameters of the three-parameter model when compared to the NDIR procedure. However, as stated in Green et

al. (1982), "...when considering a new procedure, the sample size requirements must be reevaluated using both simulation and live data... Several different sample sizes should be produced so that the effect on calibration can be determined" (p. 41).

As mentioned, this study was done in strict accordance with normality assumptions. The effects of non-normality may vary from application to application. The ease of use and cost effectiveness of the procedure make it possible to do simulation studies for many different possible applications. Researchers can easily set up simulation studies that represent their individual applications. When severely distorted distributions occur, further adjustments can be made. For example, if a distribution is skewed due to missing cases below or above a given ability range, the midpercentile values could be adjusted to take this into account. Routines for omitted and not reached items could also be written into the program. Percentage scores could even be used to compute the z' scores.

References

Ashler, D. (1979). Biserial estimators in the presence of guessing. Journal of Educational Statistics, 4, 325-355.

Baker, F. B. (1987). Methodology review: item parameter estimation under the one-, two-, and three-parameter logistic models. Applied Psychological Measurement, 11, 11-141.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26, 4, 347-372.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1982). Evaluation plan for the computerized adaptive vocational aptitude battery. (Research Report 82-1). Baltimore, Md: The Johns Hopkins University Department of Psychology.

Gugel, J. F. (1988). Evaluation of a normalized direct approach to estimating the parameters of the normal ogive three-parameter model for ability tests. (Doctoral dissertation, University of Maryland). University Microfilms International. 8808556.

Guilford, J. P. & Fruchter, B. (1978). Fundamental statistics in psychology and education (6th Ed.). NY: Mc Graw Hill.

Guilford, J. P. (1954). Psychometric methods. NY: Mc Graw Hill.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A

monte carlo study. Applied Psychological Measurement, 6, 249-260.

IBM Corp. (1968). System/360 Scientific Subroutine Package. P360-CM-03X) Version III. White Plains, NY: Technical Publications Department. 78, 83-84.

Johnson, A. P. (1951). Notes on a suggested index of item validity: The U-L index. Journal of Educational Psychology, 42, 499-504.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30, 17-24.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lord, F. M. (1980) Applications of item response theory to practical testing problems. NJ: Lawrence Erlbaum Associates.

Morrison, D. F. (1967) Multivariate statistical methods. NY: McGraw Hill.

Rowley, G. L., & Traub, R. F. (1977). Formula scoring, number-right scoring, and test taking strategy. Journal of Educational Measurement, 14, 15-22.

Schmidt, F. L. (1977). The Urry method of approximating item parameters of latent trait theory. Educational and Psychological

Measurement, 37, 613-620.

Urry, V. W. (1976). Ancillary estimators for the item parameters of mental tests. In W. A. Gorham (Chair), Computerized testing: Steps toward the inevitable conquest. (PS-76-1). Washington, D. C.: U. S. Civil Service Commission, Personnel Research and Development Center.

Urry, V. W. (1977). Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 14, 181-196.

Wingersky, M. S., Barton, M. H.,& Lord, F. M. (1983). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. LOGIST 5, Version 1. Princeton, NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (Research Memorandum 76-6). (Revised). Princeton, NJ: Educational Testing Service.

Figure 1

Two Bivariate Surface Slices (3 & 4) With Ability (1) and Item (2) Marginal (Total Group) Distributions

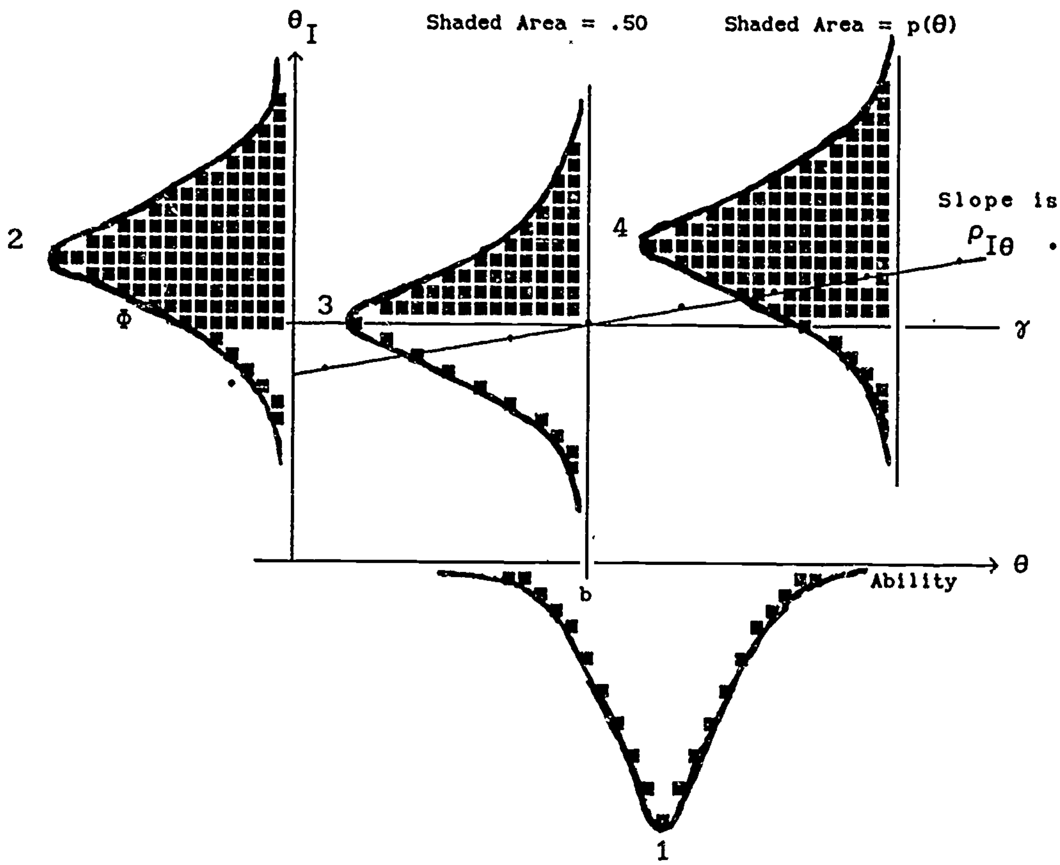

TABLE 1

Three Synthetic Test Administrations With Raw Scores (X) Associated With Underlying θ, Their Expected Freq. Prop., θ Mean, z/X, and z'/X for a set of Item Parameters That Vary Only on the Guessing Parameter

	No Guessing				Guessing = .10				Guessing = .20			
X	Freq	θ Mn	z/X	z'/X	Freq	θ Mn	z/X	z'/X	Freq	θ Mn	z/X	z'/Z
0	.002	-2.85	-2.43	-3.13	.000	---	---	---	.000	---	---	---
1	.004	-2.56	-2.31	-2.68	.000	---	---	---	.000	---	---	---
2	.006	-2.32	-2.18	-2.39	.001	-2.64	-2.66	-3.28	.000	---	---	---
3	.008	-2.12	-2.06	-2.17	.002	-2.51	-2.52	-2.95	.000	---	---	---
4	.009	-1.94	-1.94	-1.99	.003	-2.38	-2.39	-2.67	.000	---	---	---
5	.012	-1.78	-1.82	-1.83	.005	-2.23	-2.26	-2.43	.001	-2.48	-2.78	-3.21
6	.014	-1.64	-1.70	-1.68	.007	-2.08	-2.12	-2.22	.001	-2.38	-2.64	-2.94
7	.016	-1.50	-1.58	-1.54	.009	-1.93	-1.99	-2.03	.002	-2.27	-2.49	-2.70
8	.019	-1.37	-1.46	-1.41	.012	-1.78	-1.86	-1.86	.004	-2.16	-2.34	-2.47
9	.022	-1.25	-1.34	-1.28	.014	-1.64	-1.73	-1.70	.006	-2.03	-2.20	-2.25
10	.025	-1.13	-1.21	-1.16	.017	-1.49	-1.59	-1.55	.009	-1.90	-2.05	-2.07
11	.028	-1.01	-1.09	-1.04	.020	-1.36	-1.46	-1.41	.012	-1.76	-1.90	-1.90
12	.031	-.90	-.97	-.92	.024	-1.23	-1.33	-1.27	.015	-1.62	-1.76	-1.73
13	.034	-.78	-.85	-.80	.027	-1.10	-1.20	-1.14	.018	-1.48	-1.61	-1.57
14	.036	-.67	-.73	-.68	.031	-.97	-1.06	-1.01	.022	-1.34	-1.46	-1.41
15	.039	-.55	-.61	-.57	.035	-.85	-.93	-.88	.026	-1.20	-1.32	-1.26
16	.041	-.44	-.49	-.45	.038	-.73	-.80	-.75	.030	-1.06	-1.17	-1.12
17	.043	-.33	-.36	-.34	.041	-.61	-.66	-.63	.035	-.93	-1.02	-.98
18	.044	-.22	-.24	-.23	.044	-.49	-.53	-.50	.039	-.80	-.88	-.84
19	.045	-.11	-.12	-.11	.046	-.37	-.40	-.38	.043	-.67	-.73	-.70
20	.045	.00	.00	.00	.048	-.25	-.26	-.25	.047	-.54	-.59	-.56
21	.045	.11	.12	.11	.049	-.13	-.13	-.13	.050	-.41	-.44	-.42
22	.044	.22	.24	.23	.050	-.01	-.00	-.01	.052	-.28	-.29	-.29
23	.043	.33	.36	.34	.049	.11	.13	.12	.054	-.15	-.15	-.15
24	.041	.44	.49	.45	.048	.23	.26	.24	.055	-.02	.00	-.01
25	.039	.55	.61	.57	.047	.35	.40	.37	.055	.11	.15	.12
26	.036	.67	.73	.68	.045	.47	.53	.49	.054	.24	.29	.26
27	.034	.78	.85	.80	.042	.59	.66	.62	.051	.37	.44	.40
28	.031	.90	.97	.92	.039	.72	.80	.75	.048	.51	.59	.54
29	.028	1.01	1.09	1.04	.035	.84	.93	.88	.045	.64	.73	.68
30	.025	1.13	1.21	1.16	.032	.97	1.06	1.01	.041	.78	.88	.83
31	.022	1.25	1.34	1.28	.028	1.10	1.20	1.14	.037	.92	1.02	.97
32	.019	1.37	1.46	1.41	.024	1.23	1.33	1.28	.032	1.06	1.17	1.12
33	.016	1.50	1.58	1.54	.021	1.37	1.46	1.42	.028	1.21	1.32	1.28
34	.014	1.64	1.70	1.68	.018	1.52	1.59	1.57	.023	1.37	1.46	1.44
35	.012	1.78	1.82	1.83	.015	1.67	1.73	1.72	.019	1.53	1.61	1.60
36	.009	1.94	1.94	1.99	.012	1.83	1.86	1.90	.015	1.71	1.76	1.78
37	.008	2.12	2.06	2.17	.009	2.02	1.99	2.09	.012	1.90	1.90	1.99
38	.006	2.32	2.18	2.39	.007	2.23	2.12	2.31	.009	2.13	2.05	2.22
39	.004	2.56	2.31	2.68	.005	2.48	2.26	2.61	.006	2.39	2.20	2.53
40	.002	2.85	2.43	3.13	.002	2.78	2.39	3.06	.003	2.71	2.34	2.99
Mean			20.00				22.00				24.00	
S. D.			8.24				7.53				6.83	

Table 2

Discrepancies of z/X and z'/X Values in Table 1 From the Corresponding θ Means

	No Guessing			Guessing = .10			Guessing = .20		
X	Freq	z Diff	z' Diff	Freq	z Diff	z' Diff	Freq	z Diff	z' Diff
0	.002	.42	-.28	.000	---	---	.000	---	---
1	.004	.25	-.12	.000	---	---	.000	---	---
2	.006	.14	-.07	.001	-.02	-.64	.000	---	---
3	.008	.06	-.05	.002	-.01	-.44	.000	---	---
4	.009	.00	-.05	.003	-.01	-.29	.000	---	---
5	.012	-.04	-.05	.005	-.03	-.20	.001	-.30	-.73
6	.014	-.06	-.04	.007	-.04	-.14	.001	-.26	-.56
7	.016	-.08	-.04	.009	-.06	-.10	.002	-.23	-.43
8	.019	-.09	-.04	.012	-.08	-.08	.004	-.18	-.31
9	.022	-.09	-.03	.014	-.09	-.06	.006	-.17	-.22
10	.025	-.08	-.03	.017	-.10	-.06	.009	-.15	-.17
11	.028	-.08	-.03	.020	-.10	-.05	.012	-.14	-.14
12	.031	-.07	-.02	.024	-.10	-.04	.015	-.14	-.11
13	.034	-.07	-.02	.027	-.04	-.04	.018	-.13	-.09
14	.036	-.06	-.01	.031	-.09	-.04	.022	-.12	-.07
15	.039	-.06	-.02	.035	-.08	-.03	.026	-.12	-.06
16	.041	-.05	-.01	.038	-.07	-.02	.030	-.11	-.06
17	.043	-.03	-.01	.041	-.05	-.02	.035	-.09	-.05
18	.044	-.02	-.01	.044	-.04	-.01	.039	-.08	-.04
19	.045	-.01	.00	.046	-.03	-.01	.043	-.06	-.03
20	.045	.00	.00	.048	-.01	.00	.047	-.05	-.02
21	.045	.01	.00	.049	.00	.00	.050	-.03	-.01
22	.044	.02	.01	.050	.01	.00	.052	-.01	-.01
23	.043	.03	.01	.049	.02	.01	.054	.00	.00
24	.041	.05	.01	.048	.03	.01	.055	.02	.01
25	.039	.06	.02	.047	.05	.02	.055	.04	.01
26	.036	.06	.01	.045	.06	.02	.054	.05	.02
27	.034	.07	.02	.042	.07	.03	.051	.07	.03
28	.031	.07	.02	.039	.08	.03	.048	.08	.03
29	.028	.08	.03	.035	.09	.04	.045	.09	.04
30	.025	.08	.03	.032	.09	.04	.041	.10	.05
31	.022	.09	.03	.028	.10	.04	.037	.10	.05
32	.019	.09	.04	.024	.10	.05	.032	.11	.06
33	.016	.08	.04	.021	.09	.05	.028	.11	.07
34	.014	.06	.04	.018	.07	.05	.023	.09	.07
35	.012	.04	.05	.015	.06	.05	.019	.08	.07
36	.009	.00	.05	.012	.03	.07	.015	.05	.07
37	.008	-.06	.05	.009	-.03	.07	.012	.00	.09
38	.006	-.14	.05	.007	-.11	.08	.009	-.08	.09
39	.004	-.25	.1[illegible]	.005	-.22	.13	.006	-.19	.14
40	.002	-.42	.28	.002	-.39	.28	.003	-.37	.28
Mean AD		.0014	.0005		.0015	.0008		.0020	.0012

TABLE 3

ICC Recovery RMSE Values

Cases	Items	Run	NDIR Mean	NDIR S. D.	LOGIST Mean	LOGIST S. D.
1,000	40	1	.032	(.021)	.028	(.019)
		2	.029	(.013)	.038	(.023)
		3	.029	(.016)	.030	(.016)
		Mean	.030	(.016)	.032	(.019)
2,000	40	1	.021	(.011)	.030	(.020)
		2	.023	(.011)	.032	(.026)
		3	.023	(.016)	.030	(.020)
		Mean	.022	(.013)	.031	(.022)
3,000	40	1	.020	(.010)	.033	(.019)
		2	.024	(.014)	.030	(.017)
		3	.024	(.017)	.030	(.021)
		Mean	.023	(.014)	.031	(.019)
1,000	80	1	.028	(.013)	.034	(.021)
		2	.028	(.016)	.031	(.022)
		3	.025	(.016)	.033	(.021)
		Mean	.027	(.015)	.033	(.021)
2,000	80	1	.022	(.009)	.029	(.018)
		2	.018	(.009)	.027	(.017)
		3	.020	(.013)	.023	(.013)
		Mean	.020	(.010)	.026	(.016)
3,000	80	1	.020	(.011)	.022	(.012)
		2	.018	(.010)	.022	(.014)
		3	.018	(.010)	.024	(.016)
		Mean	.019	(.010)	.023	(.014)

Table 4

Item Parameter Recovery Rmsep Values

Cases	It	Run	NDIR a	NDIR b	NDIR c	LOGIST a	LOGIST b	LOGIST c
1,000	40	1	.285	.186	.070	.266	.204	.070
		2	.203	.161	.060	.272	.233	.091
		3	.228	.148	.058	.213	.178	.080
		Mean	.239	.165	.063	.250	.205	.080
2,000	40	1	.178	.150	.071	.227	.161	.082
		2	.208	.105	.061	.246	.186	.070
		3	.206	.173	.064	.211	.196	.088
		Mean	.197	.143	.065	.228	.181	.080
3,000	40	1	.163	.148	.068	.226	.194	.099
		2	.212	.136	.070	.170	.151	.078
		3	.227	.188	.078	.242	.188	.098
		Mean	.201	.157	.072	.213	.178	.092
1,000	80	1	.199	.155	.074	.229	.178	.089
		2	.197	.174	.067	.198	.188	.066
		3	.214	.142	.065	.219	.166	.087
		Mean	.203	.157	.069	.215	.177	.081
2,000	80	1	.176	.127	.053	.124	.126	.060
		2	.155	.124	.047	.180	.178	.074
		3	.166	.140	.057	.178	.160	.074
		Mean	.166	.130	.052	.161	.155	.069
3,000	80	1	.151	.134	.055	.156	.160	.074
		2	.148	.131	.063	.152	.128	.063
		3	.147	.129	.056	.123	.151	.070
		Mean	.149	.131	.058	.144	.146	.069
Grand Mean			.192	.147	.063	.202	.174	.078

Table 5

Item Parameter Correlations of Known vs Estimated Parameters

Cases	It	Run	NDIR a	NDIR b	NDIR c	LOGIST a	LOGIST b	LOGIST c
1,000	40	1	.859	.990	.554	.917	.990	.623
		2	.929	.993	.620	.910	.934	.380
		3	.923	.994	.623	.937	.992	.384
		Mean	.908	.992	.600	.922	.988	.471
2,000	40	1	.959	.993	.478	.940	.992	.496
		2	.940	.998	.706	.925	.989	.605
		3	.941	.992	.574	.950	.989	.444
		Mean	.947	.994	.594	.938	.990	.518
3,000	40	1	.971	.994	.549	.945	.990	.391
		2	.949	.995	.537	.972	.993	.529
		3	.934	.990	.379	.948	.991	.407
		Mean	.954	.993	.492	.956	.991	.444
1,000	80	1	.930	.994	.506	.924	.991	.429
		2	.934	.992	.631	.937	.990	.640
		3	.916	.994	.575	.926	.992	.425
		Mean	.926	.993	.572	.929	.990	.506
2,000	80	1	.954	.997	.728	.976	.996	.629
		2	.962	.996	.742	.956	.991	.496
		3	.950	.994	.666	.960	.994	.499
		Mean	.956	.996	.714	.965	.994	.544
3,000	80	1	.965	.995	.667	.969	.994	.548
		2	.969	.995	.609	.970	.996	.661
		3	.974	.995	.670	.978	.993	.556
		*Mean	.969	.995	.650	.972	.994	.590
*Grand Mean			.948	.994	.608	.950	.992	.514

* Weighted means obtained using Fisher's r to z transformation

Table 6

Cpu Time in Seconds Per Run

Cases	It	Run	NDIR Seconds	LOGIST Seconds
1,000	40	1	5.964	71.663
		2	5.972	77.663
		3	6.045	75.914
2,000	40	1	6.938	122.588
		2	7.059	125.921
		3	6.958	123.305
3,000	40	1	7.846	171.281
		2	7.973	174.189
		3	7.974	169.171
1,000	80	1	19.924	166.702
		2	20.008	157.851
		3	19.769	146.628
2,000	80	1	22.375	238.114
		2	22.189	252.827
		3	21.384	253.089
3,000	80	1	24.461	338.804
		2	23.987	343.425
		3	24.301	320.434